

**Concentration Inequalities**  
**Prof. Aditya Gopalan**  
**Prof. Himanshu Tyagi**  
**Department of Electrical Communication Engineering**  
**Indian Institute of Science, Bengaluru**

**Lecture - 16**  
**Concentration bounds for functions beyond bounded difference using transportation method**

(Refer Slide Time: 00:20)

Lecture 15  
24 November 2020 14:47

BEYOND BOUNDED DIFFERENCES (FEAT: THE TRANSPORTATION METHOD)

\* Review: Core ingredient of the transportation method for proving tail concentration: TRANSPORTATION LEMMA

$$\psi_Z(\lambda) \leq \frac{\lambda^2 V}{2} \quad \forall \lambda \geq 0 \quad \text{iff}$$
$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \leq \sqrt{2V D(Q||P)} \quad \forall Q \ll P.$$

\* McDiarmid's inequality using this recipe:

Reduces to exhibiting a coupling TP b/w  $X_1^n \sim P_1 \times \dots \times P_n$   
&  $Y^n \sim Q$  for which

Hi all, the aim of today's lecture is to use the transportation method which we have covered in the past two lectures to try to prove concentration inequalities for functions that do not just have the standard bounded differences property for which McDiarmid's inequality applies, but the aim is to actually relax bounded differences conditions as much as possible and in today's lecture we will actually see a very powerful example of such a condition for which the transportation method can readily give us answers.

So, before we start let us just take a few moments to review what we have seen about the transportation method.

Now the core ingredient as you recall of the transportation method or transportation cost or optimal transportation method for proving tail concentration inequalities is what is called the transportation lemma which is essentially a bridge between control of the moment generating function or the log moment generating function. And control of

differences of expectations of the quantity that we want to control with respect to different probability measures.

So, the transportation lemma basically says one version of the transportation lemma says that you can enjoy sub Gaussian moment generating log moment generating functions if and only if you can show that certain transportation cost type terms like  $E Q Z - E P Z$  for do different probability measures  $Q$  and  $P$  can be bounded let us say as square root of the KL divergence between them, so these two statements are equivalent.

(Refer Slide Time: 02:06)

Reduces to exhibiting a coupling TP b/w  $X_1^n \sim P_1 \times \dots \times P_n$   
 &  $Y_1^n \sim Q$  for which

$$\sum_{i=1}^n \mathbb{P}(X_i \neq Y_i)^2 \leq \frac{1}{2} D(Q \| P).$$

This is exactly MARTON'S transp. cost lemma

- ↳ Pinsker's ineq. +
- ↳ Maximal coupling +
- ↳ Chain rule for KL divergence

And in the last lecture we were able to prove the famous McDiarmid's inequality or bounded differences inequality using this very recipe which is given by the transportation lemma. So, in that in the course of doing that we essentially reduce the problem of showing McDiarmids inequality for a bounded differences function  $f$  or  $Z = f(x)$ .

By exhibiting by reducing to the problem of exhibiting a coupling between a product measure which is the distribution of the  $X$ 's on which the function  $f$  is applied and the other measure to be coupled is an arbitrary measure  $Q$  over  $n$  symbols let us call them  $Y_1$  through  $n$  such that the following inequality held which is that the summation of the squares of the discrepancies at each coordinate should be bounded by the KL divergence ok, the overall KL divergence between the enfold probability measures.

And this is basically like an optimal transportation cost inequality which is called Marton's transportation cost which Marton's transportation cost lemma actually shows. So, this is exactly the content of Marton's transportation cost lemma that we proved earlier. The ingredients in showing this result will basically Pinsker's inequality that helped us deal with the  $n = 1$  case.

So, in  $n = 1$  case it says that the probability of  $X_i \neq Y_i$  should be at most the square root of half the KL divergence between  $x$  and  $y$  between the distributions of  $X$  and  $Y$  and this is exactly the famous Pinsker's inequality which was proved using the notion of a maximal coupling ok. And to extend it from  $n = 1$  to general  $n$  we use an induction argument at the heart of which was the chain rule for KL divergence.

(Refer Slide Time: 04:18)

This is exactly MARTON'S transp. cost lemma

- ↳ Pinsker's ineq. +
- ↳ Maximal coupling +
- ↳ Chain rule for KL divergence:

$$D(Q_{Y^n} || P_{X^n}) = \sum_{i=1}^n D(Q_{X_i|Y^{i-1}} || P_{X_i|X^{i-1}} | Q_{Y^{i-1}})$$

$$E_{Y^{i-1} \sim Q_{Y^{i-1}}} [D(Q_{X_i|Y^{i-1}} || P_{X_i|X^{i-1}} | Q_{Y^{i-1}})]$$

So, just to remind you of what this chain rule looks like it basically says that the divergence between  $Q$  on  $n$  symbols and the product measure  $P$  on  $n$  symbols or in fact, any measure  $P$  on  $n$  symbols is = the sum over all  $i$  of the conditional KL divergence of  $Y_i$  given  $Y$  all the way from 1 to  $i - 1$  with respect to the  $P$  measure or the conditional  $P$  measure for  $X_i$  given all symbols  $X$  from 1 to  $i - 1$ .

Recall that the notation  $X$  subscript something superscript something denotes the vector consisting of  $X$  starting from the lower index and all the way moving up to the upper index and when convenient we will omit the 1 in the subscript as a default ok. So, this is

conditioned on the realization  $Q$  realization of  $Y_{1:i-1}$  and averaged out with respect to the  $Q$  measure of  $Y_1$  through  $i-1$  ok.

So, just to make it a little more precise this is this term is just the conditional KL divergence is just defined as expected value of a vector of  $i-1$  symbols distributed according to  $Q$  the marginal of  $Q$  on the first  $i-1$  symbols. So, once you have fixed  $y_1$  to  $y_{i-1}$  you just go ahead and evaluate the standard KL divergence of the conditional measures.

$Y_{1:i-1} = \text{small } y_{1:i-1}$  with respect to the same thing done for  $P$   $Y_i$  given  $Y_{1:i-1} = \text{small } y_{1:i-1}$ . So, this is the expression for the chain rule for KL divergence and we saw that you could essentially extend a coupling for  $n-1$  symbols  $X$   $Y$  up to length  $n-1$  using coupling ideas to a new coupling which covers an extra symbol.

So,  $X_n$  coupling to  $Y_n$  ok. So, that was basically a recap of what we did with the transportation method in order to prove the bounded differences inequality or McDiarmid, McDiarmid's inequality.

(Refer Slide Time: 06:52)

E2 207 Concentration Inequalities 2020 - Lectures

\* BEYOND BOUNDED DIFFS (MCDIARMID) USING TRANSPORTATION

Consider  $f: \mathcal{X}^n \rightarrow \mathbb{R}$  such that

(\*) 
$$f(x) - f(y) \leq \sum_{i=1}^n c_i(x) \mathbb{1}_{\{x_i \neq y_i\}} \quad \forall x, y \in \mathcal{X}^n.$$

let  $Z := f(X) = f(X^n)$ , where  $\{X_i\}_{i=1}^n$  are indep.

Now, let us move let us try to move beyond this and try to see how much flexibility the transportation method allows us to deal with functions that are not strictly having bounded difference properties. So, let us call this beyond bounded differences ok which in other words is also McDiarmid McDiarmid's inequality using transportation using ideas from transportation.

Now let us consider a setting as follows; consider a function  $f$  which is not which satisfies a property that is weaker than the bounded differences property. So,  $f$  is a function such that let us say  $f(x) - f(y)$  can be bounded by an  $x$  dependent type bounded difference vector of the form  $c_i$  if  $x_i$  and  $y_i$  differ, for all vectors  $x$  and  $y$  in the Cartesian product  $\mathcal{X}$ . So, this property let us call this property for future use as property star.

So, note that if  $c_i$  of  $x$  if the  $i$ th coefficient as a function of the entire  $x$  the vector  $x$  does not depend on  $x$ , but it is just some constant  $c_i$ . Then this is exactly the bounded differences property satisfied with the vector  $c_1$  up to  $c_n$ , but; however, we have been more flexible in that the coefficient the sensitivity at the  $i$ th coordinate if there is a difference depends on  $c_i$  of  $x$  and that  $c_i$  can depend on  $x$  ok.

So, this is strictly more general than the standard bounded differences property and we will also assume that  $Z$  is the result of applying  $f$  to  $x$  to the vector  $x$ .

So,  $X$  is really  $X_1$  through  $n$  or omitting the subscript where the  $X_i$  are all independent ok. So, this is the setting and will actually show that the transportation method actually kicks in very neatly to help us bound the tails of set. So, before that let us recall that we have actually considered concentration of such functions beyond bounded differences in our application of the entropy method that we saw before the transportation method.

(Refer Slide Time: 09:47)

Recall a typical "beyond Bounded diffs." result we derived using the ENTROPY method earlier:

$$Z = f(X)$$

$$Z_i(x) \equiv Z_i := \inf_{x_i'} f(x^{i-1}, x_i', x_{i+1}^n) \quad \forall i \in [n].$$

If  $\sum_{i=1}^n (Z - Z_i)^2 \leq v$  (a.s.), then

$$\forall t \geq 0 \quad P[Z \geq \mathbb{E}Z + t] \leq e^{-\frac{t^2}{2v}},$$

using a "modified log Sobolev inequality".

So, let us recall that before we move on recall a typical let us say “beyond bounded differences” result let me derive using the Entropy method ok earlier. So in fact, the results from the entropy method do help us analyze this situation above. So, if  $Z = f(x)$  and  $Z_i$  is defined as the minimum over the  $i$ th coordinate of  $f$  when you fix all the other components to be the components of  $x$ .

So, this is  $Z_i$  this is really  $Z_i$  as a function of  $x$  ok. In fact, yeah. So,  $x_{i-1}$ ,  $x_i$  prime and  $x_{i+1}$  going all the way up to  $n$  for all  $i$  and if we have that the summation of  $i = 1$  to  $n$  the difference between  $Z$  and  $Z_i$  square is bounded above by an absolute constant  $v$  it is almost surely.

Then the entropy method helped us establish an upper tail inequality saying that for all  $t$  positive the probability under the natural measure of  $X_1$  through  $X_n$  of  $Z$  exceeding its own expectation  $+t$  is at most let us say  $e^{-t^2/2v}$ . So, we got a sub Gaussian tail bound with the sub Gaussianity constant depending on this upper bound on the almost sure upper bound on summation  $Z - Z_i$  whole square ok.

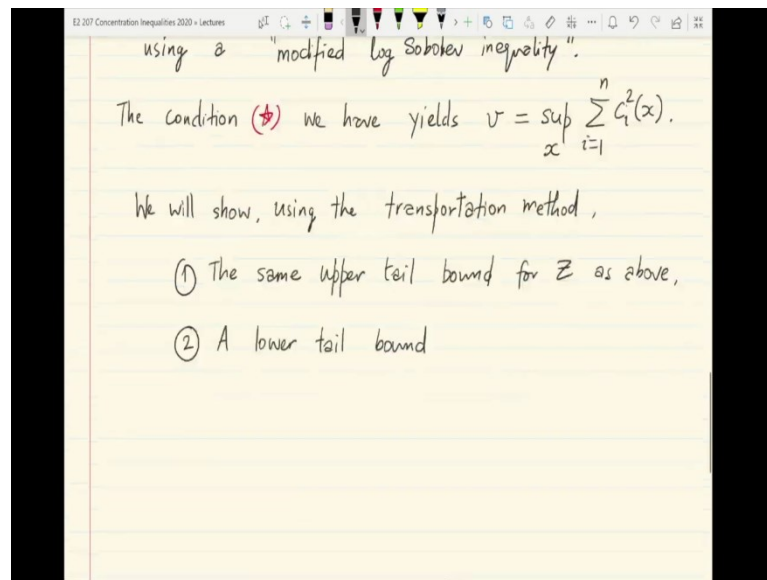
So, this was basically the way we did it one of the key ingredients in this proof using the entropy method was to first show a “modified log Sobolev inequality” if you recall ok and that helped us basically apply the entropy method to derive such a result..

So, how can we map this to the problem setting at hand? So, if you recall so, we need to satisfy this condition here ok. And if we have the condition star for our function  $f$  you can easily argue that an upper bound  $v$  for  $Z - Z_i$  whole square is can be used with the summation  $c_i x$  the whole square maximum over all possible vectors  $x$  ok pardon me. So, this is basically because if you have condition star then the least that  $f(y)$  can get ok.

So, if you just think of holding  $x$  constant at all positions, but position  $i$  and minimizing and trying to find the least value of  $y$  where  $y$  is defined as  $x$  in all, but the  $i$ th coordinate and letting it range freely over the  $i$ th coordinate then on the right hand side you basically have only one term which is the indicator  $x_i \neq y_i$  and for that term and almost sure upper bound is  $c_i$  of  $x$  ok.

So, by analogy you can just sum over all  $n$  coordinates of the squares of the differences to get  $v$  as to get summation  $c_i$  square  $x$  as a proxy for  $v$ .

(Refer Slide Time: 13:57)



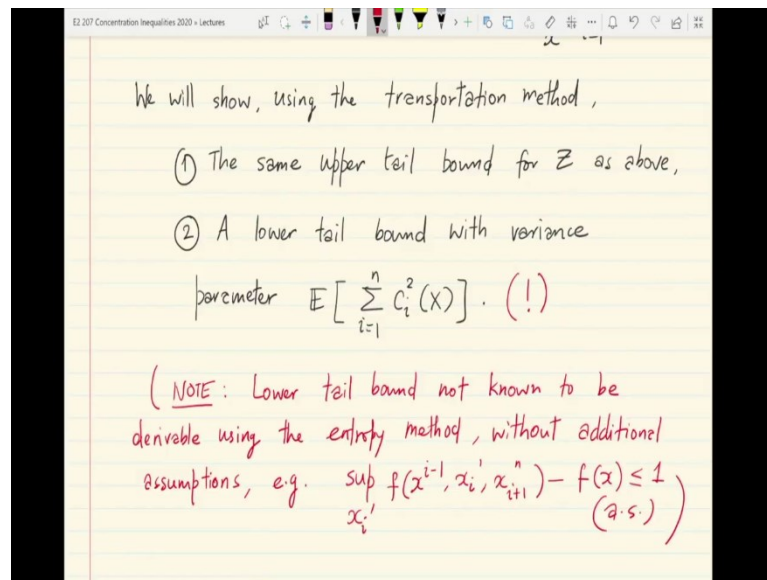
So, let me just remark here that the condition star that we have yields  $v$  = the supremum over all pairs  $x$  of the sum are  $= 1$  to  $n$   $c_i$  square of  $x$  ok. So, you can use you can apply the entropy method to derive this inequality with  $v$  being exactly this quantity here the maximum over all  $x$  of  $c_i$  square  $x$ .

Now there are several unanswered questions among them is the issue of what about the tail bound for the other side. So, this is only a right side tail bound, what about the left tail and so on. So, in fact what we will show now using the transportation method are two things. So, we will show using the transportation method, two things one is we will recover the same upper tail bound here using ideas purely from transportation optimal transportation.

So, the same upper tail bound for  $Z$  as above and moreover we will rather remarkably show without any extra effort a lower tail bound as well. So, a lower tail bound that is the probability that  $Z$  is less than  $= E Z - t$  with much better variance parameter in fact, with variance parameter of only the expectation of the sum of  $c_i$  square  $X$  ok.

Not even the supremum value of  $c_i$  square  $X$  overall  $x$  supremum of summation  $c_i$  square  $X$ , but in fact, a potentially much smaller parameter which is only the average value of summation  $c_i$  square  $X$  ok without any extra effort ok. So, this is what the transportation method allows us to show.

(Refer Slide Time: 16:34)



As a side note if you are wondering what about the scope of the entropy method to prove a lower tail bound.

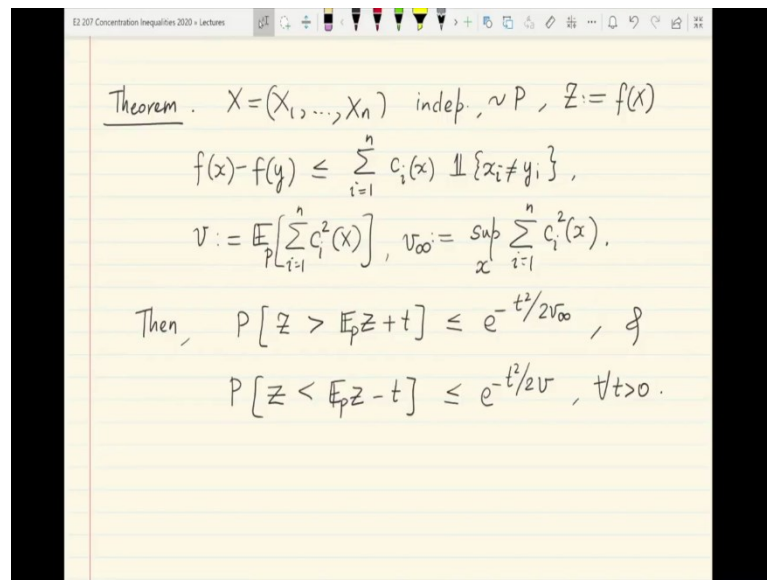
In fact, there is no such lower tail bound known using ideas from the entropy method without additional assumptions. So, the lower tail bound not known currently not currently known to be drivable using the entropy method, alone without perhaps additional assumptions that have to be made. So, one example of an assumption that gives you a lower tail bound via the entropy method which is there in the textbook the concentration inequality textbook is of the following form that if you consider.

So, if you consider yeah if you consider maximizing  $f$  at the  $i$ th location when all else is fixed  $x_{i+1}$  to  $n$  ok and take the difference of this with the original  $f(x)$  then for instance this must be bounded by 1 almost sure. So, such an assumption helps you prove lower tail bounds, but even then the lower tail bound is does not have this improved variance parameter, you can look this lower tail bound derivation up in the textbook during the entropy method.

But what is remarkable rather remarkable here is that the transportation method easily allows us to prove these two results with essentially the same amount of effort ok. So, let us go ahead and write down the formal result which we will argue in this and the next lecture using the transportation method. So, this is the following result.



(Refer Slide Time: 18:38)



Theorem .  $X = (X_1, \dots, X_n)$  indep.  $\sim P$ ,  $Z := f(X)$   
 $f(x) - f(y) \leq \sum_{i=1}^n c_i(x) \mathbb{1}_{\{x_i \neq y_i\}}$ ,  
 $v := \mathbb{E}_P \left[ \sum_{i=1}^n c_i^2(X) \right]$ ,  $v_{\infty} := \sup_x \sum_{i=1}^n c_i^2(x)$ .  
 Then,  $P[Z > \mathbb{E}_P Z + t] \leq e^{-t^2/2v_{\infty}}$ , &  
 $P[Z < \mathbb{E}_P Z - t] \leq e^{-t^2/2v}$ ,  $\forall t > 0$ .

So, let us assume  $X$  is a vector of independent random variables are distributed according to the probability distribution  $P$  which is a product measure over  $n$  coordinates and let us  $Z = f(x)$  and we have  $f$  satisfying the property that you can bound  $f(x) - f(y)$  in terms of  $x$  dependent hamming sensitivities.

And let us also denote  $v$  as the expected value under the natural distribution of  $x$  of summation  $c_i^2$  of  $X$  and  $v_{\infty}$  as the largest value of such a sum of squares squared coefficients. So, sup over all possible arguments  $x$  of  $\sum_{i=1}^n c_i^2(x)$ . So,  $v_{\infty}$  is clearly lower bounded by  $v$ .

And the result of the theorem is that under the  $P$  measure the measure of the independent  $x$  is the probability  $Z$  exceeds  $\mathbb{E}_P Z + t$  is at most  $e^{-t^2/2v_{\infty}}$  and on the lower tail side on the left tail side probability  $Z$  less than expected value of  $Z - t$  is at most  $e^{-t^2/2v}$  which is a much lesser sub gaussianity constant than  $v_{\infty}$  in many cases in this whole for all  $t$  positive.

So, this is the content of this beyond bounded differences kind of result for which the transportation method comes in very handy as we can see.

(Refer Slide Time: 21:13)

PROOF: Let  $Q \ll P$ , & let  $(X, Y) \sim P$  be a coupling of  $(P, Q)$ ,  $P = P_1 \times \dots \times P_n \equiv P_1^n \equiv P^n$ .

$$\begin{aligned} \mathbb{E}_Q[Z] - \mathbb{E}_P[Z] &= \mathbb{E}[f(Y) - f(X)] \\ &\leq \mathbb{E}\left[\sum_{i=1}^n c_i(Y) \mathbb{1}_{\{Y_i \neq X_i\}}\right] \quad (f \text{ smoothness property } (\star)) \\ &= \mathbb{E}\left[\sum_{i=1}^n c_i(Y) \mathbb{P}[Y_i \neq X_i | Y]\right] \quad (\text{iterated expectation}) \\ &\leq \mathbb{E}\left[\sqrt{\sum_{i=1}^n c_i(Y)^2} \cdot \sqrt{\sum_{i=1}^n \mathbb{P}[Y_i \neq X_i | Y]^2}\right] \quad (\text{Cauchy-Schwarz}) \\ &\leq \sqrt{\mathbb{E}\left[\sum_{i=1}^n c_i(Y)^2\right]} \cdot \sqrt{\mathbb{E}\left[\sum_{i=1}^n \mathbb{P}[Y_i \neq X_i | Y]^2\right]} \quad (\text{Cauchy-}) \end{aligned}$$

So, let us move on to the proof of this result which will take us some time to execute, but at a high level the I will outline in this lecture the high level idea of the proof. So, we are going to use a transportation lemma. So, the first order of business is to consider probability measure  $Q$  absolutely continuous with respect to  $P$ .

And let  $X$  comma  $Y$  distributed according to  $P$  we are coupling of the measures  $P$  and  $Q$  ok in that order. So,  $X$  is distributed to the marginal  $X$  follows the distribution  $P$  the marginal  $Y$  is distributed according to  $Q$  and  $P$  is a product measure ok. So,  $P$  is really  $P_1$  all the way up to  $P_n$  if we will also abbreviate as  $P_1$  to  $n$  or  $P_n$  occasionally.

So, let us consider bounding the expected value of  $Z$  with respect to  $Q$  - the expected value of  $Z$  with respect to  $P$  as the transportation lemma requires. So, by coupling this is just =  $\mathbb{E}$  under the coupling measure the joint distribution  $P$  of  $f(Y) - f(X)$  ok.

Now once you have  $f(Y) - f(X)$  together you can invoke the beyond bounded differences kind of property or the smoothness property for  $X$  in the in terms of the hamming distance to say that this is upper bounded by the sum of  $c_i$  of  $Y$  into indicator  $Y_i \neq X_i$ . So, this is by the  $f$  smoothness property which is also the equation star that we wrote earlier.

Now, what we can do is, we can take we can condition this expectation iteratively on  $Y$  itself the entire vector  $Y$ . So, imagine taking a double expectation with respect to  $Y$  and so if you do that you basically get the expected value of sum  $i = 1$  to  $n$   $c_i$  of  $Y$  into the

expected value of the indicator given  $Y$  that is just the probability of  $Y_i \neq X_i$  given the vector  $Y$ . So, this is just by the property of iterated expectation ok.

And now what we can do is the thing inside the expectation looks like an inner product it is the summation of  $i = 1$  to  $n$  into  $\sum a_i$  into  $\sum b_i$ . So, to upper bounded we can use something like Cauchy Schwarz whereby we get  $i = 1$  to  $n$   $c_i Y_i^2$  under the square root into another summation of all these probabilities  $Y_i \neq X_i$  continue given  $Y$  the whole square under the square root ok.

So, this is basically by Cauchy Schwarz and which is inside the expectation nothing about the expectation is being used here, at this point to continue upper bound in this we can now use the probability version of Cauchy Schwarz. So, which basically says that the expected value of a product of random variables. So, there are two random variables sitting here this is the first random variable and this is the second random variable.

The product of two random variables the expected value of two random variables is at most the  $L^2$  norm in some sense of these the product of the  $L^2$  norms of these random variables were  $L^2$  norm defined with respect to the expectation operator. So, formally this just becomes the expected value of some  $c_i Y_i^2$  under the square root into another square root of the expected value of the sum of all these probabilities conditional probabilities square ok.

This is again by Cauchy Schwarz in its probabilistic flavour ok and it useful to notice here that the quantity inside the first expectation by definition is almost surely upper bounded by the infinity ok, as per our definition of  $v_\infty$ ,  $v_\infty$  is basically the largest possible value of the sum of  $c_i$  squares ok.

(Refer Slide Time: 26:38)

$$\begin{aligned} & \leq \sqrt{\mathbb{E}\left[\sum_{i=1}^n c_i(Y)^2\right]} \cdot \sqrt{\mathbb{E}\left[\sum_{i=1}^n \mathbb{P}[Y_i \neq X_i | Y]^2\right]} \quad (\text{Cauchy-Schwarz}) \\ & \leq \sqrt{V_\infty} \cdot \sqrt{\mathbb{E}\left[\sum_{i=1}^n \mathbb{P}[X_i \neq Y_i | Y]^2\right]} \\ \text{Similarly,} \\ \mathbb{E}_Q[-Z] - \mathbb{E}_P[-Z] & \leq \sqrt{\mathbb{E}\left[\sum_{i=1}^n c_i(X)^2\right]} \cdot \sqrt{\mathbb{E}\left[\sum_{i=1}^n \mathbb{P}[X_i \neq Y_i | X]^2\right]} \\ & = \sqrt{v} \end{aligned}$$

So, finally, we get that  $\mathbb{E}_Q[f(Z)] - \mathbb{E}_P[Z]$  is upper bounded by square root  $v_\infty$  into the square root of this sort of transportation cost like them ok. So, we get this now along very similar lines we can apply the same technique to the random variable  $-f(x)$  or  $-Z$  ok to get the following.

So,  $\mathbb{E}_Q[-Z] - \mathbb{E}_P[-Z]$  can in a very similar fashion be upper bounded. So, what you will have now is, basically you will have the expected value of  $c_i$  summation  $c_i X$  square through expected value. So, we will have  $X$  instead of  $Y$  ok.

We will have  $X$  playing the role of  $Y$  into the same square root of the expected value of the sum of these conditional probabilities given  $X$  instead of given  $Y$  ok and by definition this is exactly  $= v$  because now the expectation is over the  $X$  or the  $P$  measure ok. So, notice that we have basically obtained two bounds one for the difference of expectations of  $Z$  and one for the difference of expectations of  $-Z$ . In both cases there is a constant that shows up which is either  $v_\infty$  or  $v$ .

So, the  $Z$  inequality will help us give a right tail bound for  $Z$  and the  $-Z$  inequality will help us give a left tail bound concentration bound for  $Z$  and all we need to handle is that we need to show that the second term which is this transportation like cost which is the expected value of sum of conditional probabilities of discrepancy square should somehow be related or upper bounded by the divergence between the measures ok.

(Refer Slide Time: 29:17)

$= V.$

Therefore, via the transportation lemma, we will be done if we show:

MARTON'S CONDITIONAL TRANSPORTATION COST INEQUALITY:

$$\min_{(X,Y) \in \mathcal{P}(P,Q)} \mathbb{E} \left[ \sum_{i=1}^n \mathbb{P}[X_i \neq Y_i | X]^2 + \mathbb{P}[X_i \neq Y_i | Y]^2 \right] \leq 2D(Q||P)$$

$\forall Q \ll P \equiv P_1 \times \dots \times P_n.$

So, therefore, the upshot of this calculation is that via you know via the transportation lemma we will be done if we can show ok what is called Marton's conditional transportation cost inequality which is essentially the statement that there exists a coupling of  $X$  and  $Y$  or of the measures  $P$  and  $Q$ , such that the under that coupling the expected value of the conditional probability of discrepancy the square of it given  $X$  + the same symmetric version given  $Y$  ok.

So, the sum of these two conditional discrepancy probability squares is at most  $2D(Q||P)$  for all probability measures  $Q$  absolutely continuous with respect to  $P$  which is basically assumed to be a product measure ok. So, this is exactly Marton's conditional transportation cost inequality which we will show in the next lecture.

But assuming this what one can do is one can just go and use it in the appropriate tail bounds to bound these conditional probabilities, they are bounding the sum of these conditional probabilities probability squares. So, that naturally implies the same bound for each of these conditional probability squares and you can relate it to the divergence and then use the transportation lemma to sort of transfer it to be control of the log moment generating function.

And finally, Chernoff will kick in to give you the appropriate tail bound ok. So, that basically is the summary of how the transportation method gives you both tail bounds with

the appropriate variance parameters  $v$  infinity or  $v$  just by invoking this powerful Marton's conditional transportation cost inequality which will be the subject of a next lecture.

Thank you.